In search of a core cellular network in single cell transcriptome data

Running title: A core gene network

Ming Yang[1], Benjamin R. Harrison[1], Daniel E.L. Promislow[1,2*]

[1]Department of Laboratory Medicine and Pathology, University of Washington School of Medicine, Seattle, WA 98195 USA

[2]Department of Biology, University of Washington, Seattle, WA 98195 USA

***Correspondence:**

Daniel Promislow

promislo@uw.edu

Abstract

Along with specialized functions, cells of multicellular organisms also perform essential functions common to most if not all cells. Whether diverse cells do this by using the same set of genes, interacting in a fixed coordinated fashion to execute essential functions, remains a central question in biology. Single-cell RNA-sequencing (scRNA-seq) measures gene expression of individual cells, enabling researchers to discover gene expression patterns that contribute to the diversity of cell functions. Current efforts to study cellular functions focus primarily on identifying differentially expressed genes across cells. However, patterns of co-expression between genes are probably more indicative of biological processes than are the expression of individual genes. Using single-cell transcriptome data from the fly brain, here we focus on gene co-expression to search for a core cellular network. We constructed cell type-specific gene co-expression networks using single-cell transcriptome data of brains from the fruit fly, *Drosophila melanogaster*. We detected a set of highly coordinated genes preserved across cell types in fly brains and defined this set as the core cellular network. This network was further validated through two additional datasets profiling fly brains or heads. This core is very small compared with cell type-specific gene co-expression networks and shows dense connectivity. Modules within this core are enriched for basic cellular functions, such as translation and ATP metabolic processes, and gene members of these modules have distinct evolutionary signatures. Overall, we demonstrated that a core cellular network exists in diverse cell types of fly brains and this core exhibits unique topological, structural, functional and evolutionary properties.

Introduction

Life on Earth has gone through many transitions in organizational complexity (Smith and Szathmary 1997). Among these, the evolution of multicellularity stands out as a key milestone. This transition has occurred independently multiple times across the tree of life and paved the way for tremendous phenotypic expansion and biological diversification (Parfrey and Lahr 2013). Although this has led to the evolution of cell type-specific regulatory pathways that define cells with vastly different functions, all cells in multicellular organisms also carry out common functions that are essential for cell survival. Whether these common functions are supported by a shared network of coexpressed genes, remains a central question in biology (Lim et al. 2013; Hart and Alon 2013). In particular, do all cells utilize the same set of genes in a coordinated fashion to accomplish common functions, to form a core regulatory network?

Cellular phenomena can be characterized by different endophenotypic domains or levels of biological organization, such as the genome, epigenome, transcriptome, proteome, etc. Investigating core functions from these different levels not only provides insight into essential functions of cellular life, but also helps to reveal the evolutionary forces acting at different levels of biological organization (Sorrells and Johnson 2015; Ghadie et al. 2018; Wagner 2012). To identify core functions at each level, researchers have used various strategies such as identifying genes that are expressed constitutively over temporal or spatial scales, and across environments. These genes are typically referred to as 'housekeeping genes' and are thought to perform essential functions. Housekeeping genes tend to be evolutionarily ancient (Zhu et al. 2008), highly conserved (Zhang and Li 2004), and are enriched for several functions, including metabolism,

54 RNA binding, protein degradation and cytoskeleton functions (Zhang and Li 2004; Lehner and
55 Fraser 2004). Using somewhat circular logic, core functions are often described based on
56 housekeeping genes. However, we recognize that genes do not work in isolation, but work with
57 each other to carry out biological processes. Co-expression in particular is an indicator of
58 functional relationships (Hughes et al. 2000), and so individual molecular abundances alone may
59 not adequately capture biological organization.

60 High-throughput methods that generate high-dimensional 'omic' data have greatly
61 increased our understanding of molecular and cellular function and organization, in particular
62 through the analysis of molecular networks based on co-expression (Barabasi and Oltvai 2004;
63 Proulx et al. 2005; Thompson et al. 2015; Promislow 2005). Networks consist of nodes connected
64 to one another by edges. In the search for the underlying network structure of cells, researchers
65 have explored many different kinds of edges, including but not limited to gene co-expression,
66 protein-protein interactions (PPI), interactions among transcription factors (TF), TF chromatin
67 occupancy, miRNA-target gene interactions, metabolite covariation, and metabolic reactions
68 (Mitra et al. 2013). Many studies have mainly focused on tissue-specific networks (Greene et al.
69 2015; Sonawane et al. 2017), and some have also examined networks that show some level of
70 conservation. For example, co-expression network analysis of human and *Arabidopsis* bulk
71 transcriptome data has found a substantial number of gene pairs whose co-expression spans
72 multiple datasets (Lee et al. 2004; He and Maslov 2016). In both analyses, gene pairs expressed
73 across samples were enriched for translation, DNA replication, and regulation of transcription
74 functions, all generally considered to be core cellular functions. Additionally, Skinnider et al.
75 (2021) constructed tissue-specific PPI networks using co-immunoprecipitation within each of

76 seven mouse tissues (Skinnider et al. 2021). They discovered core cellular modules, present in all
77 mouse tissues, composed of evolutionarily ancient proteins, which contrasts with evolutionarily
78 novel accessory modules that are found within individual tissues.

79 A major drawback of most previous studies investigating core functions from a network
80 perspective is that the networks were inferred from bulk data, which profiles heterogeneous cell
81 populations of an organism, organ, or tissue. Bulk samples face two main limitations for network
82 construction. First, differences in cellular compositions between samples may confound
83 covariation analysis (Farahbod and Pavlidis 2020). Second, measurements that are averaged over
84 thousands of cells in bulk samples make it difficult to detect interactions between genes in
85 individual cells, such as the presence of co-expression patterns and the cell-specificity of these
86 interactions. The compendium of housekeeping genes, initially characterized based on the
87 consistency of their expression, may need revision based on analyses of gene-gene relationships.
88 Housekeeping genes may show clear evidence of co-expression in all cell types; or alternatively,
89 interactions among housekeeping genes may be relatively weak or even cell type-specific. To
90 distinguish these possibilities requires that we define gene networks at a cellular level.

91 With the advent of single-cell RNA sequencing (scRNA-seq), we have an unprecedented
92 opportunity to reveal gene relationships in specific cellular contexts and to probe cellular-level
93 networks (Trapnell 2015; Tanay and Regev 2017). One recent study used scRNA-seq from mouse
94 brain samples to construct gene co-expression networks, comparing the topology of networks built
95 from different levels of cell type hierarchy (i.e., from broad to specific classes) (Harris et al. 2021).
96 Their results show well-preserved gene-gene relationships at each level, and suggest the existence

97 of a core co-regulatory network in the brain. However, they did not directly compare cellular
98 networks across cell types, which leaves the possibility that what appear as core networks at more
99 integrated levels may not manifest when examined by cell type, or by individual cell. Considering
100 the state of the field and the considerable amount of scRNA-seq data now available, we are led to
101 ask several fundamental questions. First, can we identify shared co-expression patterns between
102 pairs of genes across different cell types; Second, how common are these shared co-expressed gene
103 pairs among cell types? Lastly, do these shared co-expressed genes point to a core cellular network,
104 and if so, what properties does this core network manifest?

105 To investigate these questions, we used a published scRNA-seq dataset derived from whole
106 fly brains (Davie et al. 2018), and constructed cell type-specific gene co-expression networks. We
107 then looked for shared edges across cell types. We found a population of gene-pairs that are more
108 common across cell types than expected by chance, and yet, none of these edges were found in all
109 cell types. We then determined that those edges that were common among most but not all cell
110 types tended to be near, but below, the correlation threshold we used to construct gene co-
111 expression networks in the remaining cells, suggesting that the edges of a core network show
112 consistent high and varying co-expression strength over cell types. The network we uncovered is
113 small relative to the number of commonly expressed genes, which indicates that core cellular
114 functions are either more limited than previous analysis suggests, or that gene co-expression is not
115 a salient feature of functions common to all cells. We further described the core network's
116 topological properties, functional enrichment, and the evolutionary age of its constituent genes.
117 We validated our findings in two independent datasets, one based on cells from the fly brain, and
118 a second from the fly head, demonstrating the extensibility of our analytical approach. To our

knowledge, this is the first study of a core cellular network among cell types using single-cell transcriptome data in the fly brain and marks a shift in paradigm for investigating commonalities in a molecular endophenotypic domain from a network perspective at the single-cell level.

Results

Construction of cell cluster-specific gene co-expression networks

We analyzed published fly brain atlas data obtained from whole-cell fly brain samples of female and male *Drosophila* using 10X Genomics library preparation methods (Davie et al. 2018). The original dataset contained expression profiles of 17,473 genes in 56,902 high-quality brain cells. The cells were grouped into 116 cell clusters, 74 of which were annotated to known fly brain cell types (Davie et al. 2018). Here we built cell cluster-specific gene co-expression networks in females and males separately, and we investigated the consistency of covarying gene pairs across these cell clusters. We selected 68 cell clusters that contained at least 200 cells and filtered expressed genes in each cell cluster individually (**Methods**). The number of genes expressed per cell cluster ranged from 2,683 in the Tm5ab cell cluster of females, to 6,727 in cell cluster 0 of males (**Fig. S1**). In total, there were 7,795 genes expressed in at least one cell cluster, 2,088 of which were expressed in all 68 cell clusters (**Fig. S1**). scRNA-seq data are typically sparse, with the expression of many genes falling below the limits of detection. Sparsity, measured as the percentage of zeros in the data for the 2,088 commonly expressed genes, was below 50% for all

138 but one cell cluster (cluster 32 in males, **Supplementary Material, section 1)**. After excluding
139 this cell cluster from further analysis, the remaining cell clusters had less than 48% sparsity.
140 Throughout, we focus on the 2,088 commonly expressed genes to identify covarying gene pairs
141 and co-expression networks, within and across these 67 cell clusters.

142 An overview of our pipeline is shown in **Figure 1**. We used the bigScale2 algorithm
143 (Iacono et al. 2019) to compute cell cluster-specific gene correlation matrices and then built cell
144 cluster-specific gene co-expression networks by thresholding these gene correlation matrices with
145 a top percentile-based cutoff approach (**Method and Fig. S2**). While there is no objective criterion
146 for choosing a gene correlation threshold, we employed a signal-to-noise score approach, which
147 indicates the consistency of networks constructed using subsets of original data at a given
148 threshold, to guide the threshold value selection (**Supplementary Material, section 1**). The results
149 showed that most cell cluster networks had the highest signal-to-noise score when constructed at
150 the top 5% threshold (**Supplementary Material, section 1, Fig. S3 and S4)**, hence we applied
151 this cutoff and transformed gene correlation matrices into gene co-expression networks (**Fig. S5)**.
152 At this threshold we identified 1,942,694 co-expressed gene pairs (89.16% of all possible pairs
153 among 2,088 genes) that occurred in at least one of the 67 cell cluster-specific networks.

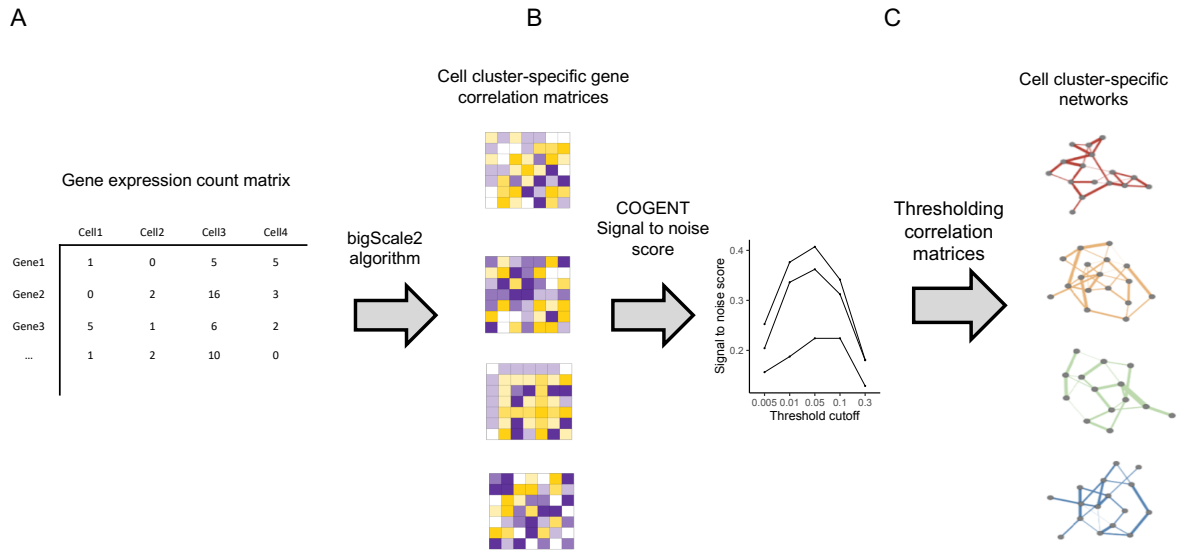

Figure 1. Pipeline overview of cell cluster-specific gene co-expression network construction. The pipeline starts with a gene expression count matrix (A) and computes a gene correlation matrix for each cell cluster using the bigScale2 algorithm (B, Iacono et al. 2019). Signal-to-noise scores are used to evaluate network robustness at different percentile thresholding values for each absolute gene correlation matrix individually, out of which a global optimal thresholding value is selected (C). Applying the selected thresholding cutoff value, gene correlation matrices are transformed into gene co-expression networks.

Co-expression networks in fly brain cell clusters are highly context-dependent

If a core cellular network exists, we expect its edges to be present in all cell clusters. We define the number of cell clusters in which each edge (i.e., each co-expressed gene pair) is detected as that edge's 'commonality'. The distribution of commonality scores was right skewed, with more than 75% of the edges specific to fewer than five cell clusters and only 0.4% of edges common to

more than 30 cell clusters (**Fig. 2A**). The largest recorded commonality score was 64, observed for only two edges. Given that 67 cell clusters were analyzed, no edges were found to be present in every cluster. As a complement to the observed edge commonality distribution, we also plotted the gene commonality distribution, where gene commonality indicates the number of cell clusters in which a given gene shared an edge with at least one other gene. The gene commonality distribution showed that most genes had one or more edges in the majority of cell clusters, and 613 genes had at least one edge in all 67 cell clusters (**Fig. 2A**). Thus, commonly expressed genes were frequently co-expressed with other genes, though the specific co-expression partners vary among different cell clusters.

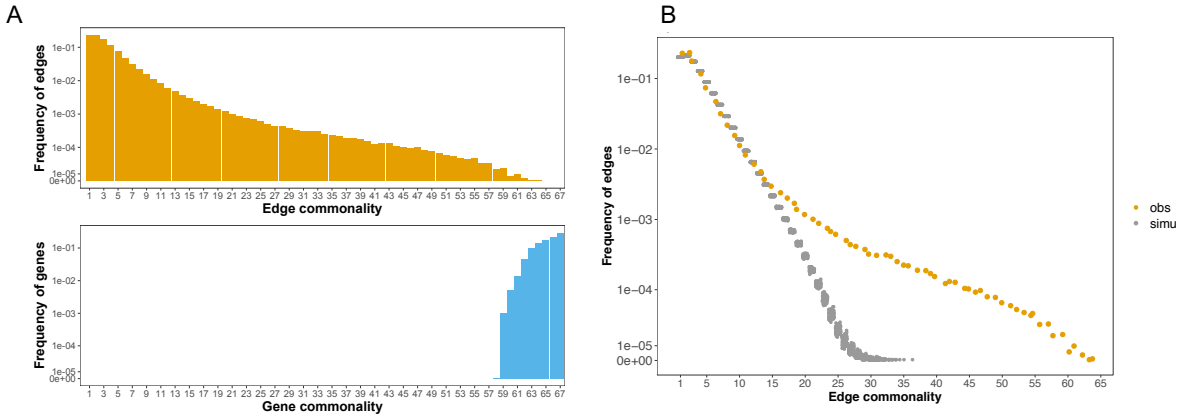

Figure 2. Cell cluster-specific co-expression networks share many more edges than random networks. A. Edge and gene commonality distributions. The commonality of an edge indicates the number of cell clusters in which that edge is detected (top). The commonality of a gene refers to the number of cell clusters in which that gene is detected as being co-expressed with one or more other one gene (bottom). The y-axis shows the frequency of genes or edges in the corresponding commonality score group. B. The observed edge commonality distribution (yellow) compared with the null expectation derived from network randomization (gray). Network randomization was performed 100

183 times for each cell cluster individually with network size (number of nodes and edges) and gene degree (number of
184 co-expressed gene partners per gene) fixed.

185

186 **Recurrently co-expressed genes in multiple cell clusters**

187 The fact that no edges were found to occur in all cells suggests either that a core cellular
188 network does not exist, or that perhaps our method for detecting co-expression was unable to
189 identify all edges of a core network. To determine if our method uncovers gene pairs that
190 recurrently co-express in multiple cell clusters, we next asked to what extent the observed edge
191 commonality distribution differed from the null expectation, where gene co-expression occurs
192 randomly among commonly expressed genes. We evaluated this in two ways. First, we derived a
193 mathematical expectation for the probabilities of edge commonality using the binomial distribution
194 (**Methods**). This calculation shows that most gene pairs were expected to co-express in only a few
195 cell clusters. For example, for a gene pair to be co-expressed in exactly 2, 3 or 4 cell clusters, the
196 probability values were 0.1970, 0.2247, or 0.1892, respectively, and the probability became
197 smaller than 0.0001 for cell cluster ≥14 (**Methods**). In a null model based on random gene co-
198 expression, the probability of one gene pair not co-expressed in any of the 67 cell types is 0.0321.
199 Thus, we would expect to find 2,108,888 unique gene pairs to occur in one or more clusters, among
200 the 2,088 commonly expressed genes (**Methods**). This number is larger than the observed
201 1,942,694 gene pairs, indicating that some genes recurrently co-express in multiple cell clusters.
202 A full comparison of this analytically predicted distribution and the observed edge commonality
203 showed that the two distributions agreed well at lower, more cell-specific commonality. However,

204 we observed a clear excess of shared edges relative to the frequency expected for those expressed
205 in 7 or more clusters (**Fig. S6**). Second, we compared the deviation between the observed edge
206 commonality distribution and a null distribution sampled using network randomization
207 (**Methods**). This comparison showed that the observed distribution was enriched for high
208 commonality edges. For instance, none of the randomizations generated an edge commonality
209 larger than 36, while the observed distribution included hundreds of such edges (**Fig. 2B**), further
210 supporting our observation that many edges occur in more cell clusters than we would expect by
211 chance. This pattern is robust to the percentile cutoff values used in network construction (**Fig.**
212 **S7**). Additionally, using a loose co-expression threshold that incorporates more edges in each cell
213 cluster network led to edge commonality distributions that more closely resembled random
214 networks (**Fig S7**). These results therefore suggest that there exists a set of co-varying genes that
215 occur more repeatedly than expected by chance across diverse cellular contexts, pointing to a core
216 cellular network composed of genes that are co-expressed across cell types.

217

218 **Rank aggregation analysis reveals a core cellular network with sub-threshold edges.**

219 An ideal core network would consist of edges that appear in every cell cluster. However,
220 in practice, the identification of gene co-expression edges based on gene expression data involves
221 the risk of false positives and false negatives. In the data sets we analyzed, no edges were present
222 in every surveyed cell cluster. This could be due to our parameter value choices, such as the
223 correlation matrix thresholding cutoff, or alternatively, these edges or gene pairs could truly be

224 cell cluster specific, and they are not co-expressed in all cell clusters. We hypothesized that if a
225 core network exists but certain edges are missing as false negatives, these edge members should
226 be consistently highly ranked among the gene pairs in all cell clusters, even if they are ranked
227 below the significance threshold in some cell types. To test this possibility, we performed a rank
228 aggregation analysis for network edges in those cell clusters where they do not pass the threshold
229 for defining an edge. Specifically, we extracted edges from a given edge commonality group *k*
230 (they belong to the top 5% in *k* cell clusters) with *k* ranging from 1 to 51, covering the spectrum
231 of edge commonalities scores, and examined each edge's ranking vector in the remaining *67-k* cell
232 clusters. The rank aggregation algorithm estimated a P value per edge ranging from 0 to 1, with a
233 small value indicating an edge is ranked consistently higher across cell clusters, and a larger value
234 meaning an edge's rank distribution over cell clusters follows a random pattern.

235 To enable fair comparison between edge commonality groups, we randomly sampled 100
236 edges in each group and computed their rank aggregation scores. The enrichment of highly ranked
237 edges indeed increases with edge commonality, a strong indication that edges in high commonality
238 groups are consistently highly ranked among the gene pairs in those cell clusters, despite being
239 below the threshold cutoff in the remaining cell clusters (**Fig. 3A**). In particular, more than 95
240 percent of edges sampled from the edge commonality groups ≥ 43 were highly ranked in the
241 remaining cell clusters ($P < 0.05$), whereas fewer than 1% of the tested edges shared by 7 or fewer
242 cell types (edge commonality ≤ 7) reached the 5% significance threshold in other clusters. Based
243 on this rank aggregation analysis and a P value < 0.05 cutoff, we chose edge commonality 43 to
244 define core network edges, which resulted in a core network in the Drosophila brain with 205
245 genes connected by 2,140 edges (**Table S1**).

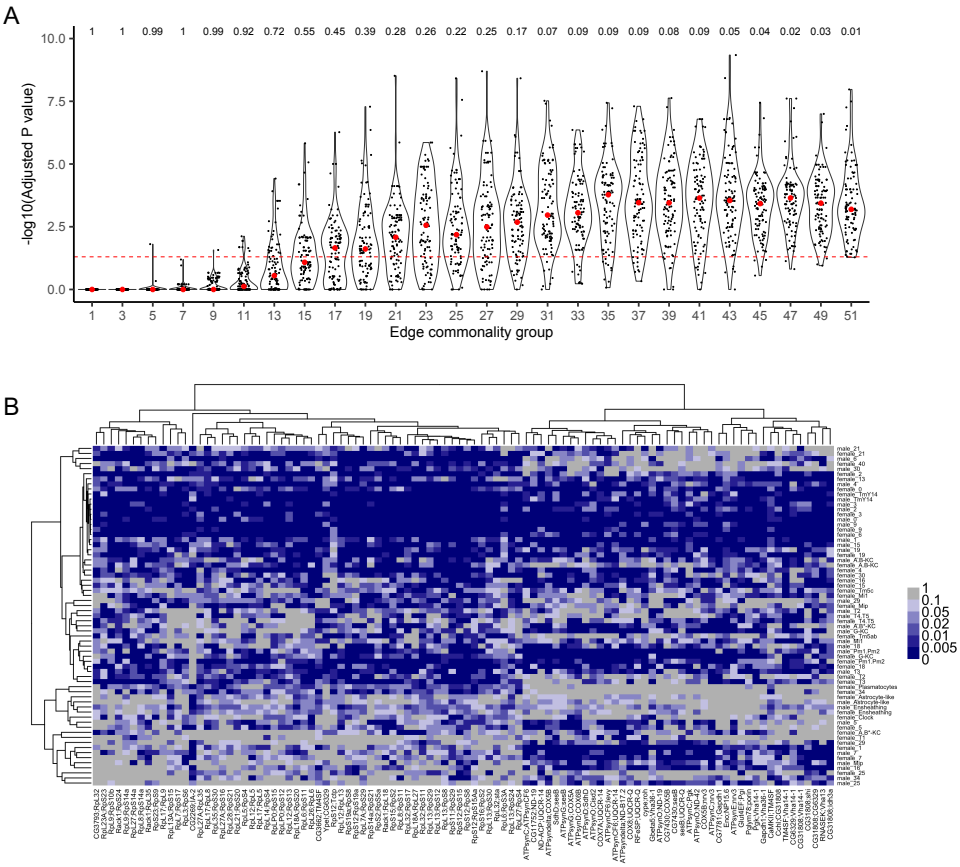

Figure 3. Identifying a core using an edge commonality cutoff based on rank aggregation analysis. A. Adjusted P-value distributions of 100 sampled edges in different edge commonality groups using a rank aggregation method. Edge commonality groups were selected to cover the full range of edge commonality scores. Each violin plot illustrates the distribution of adjusted P values for the sampled edges for each edge commonality group. Each dot represents an edge, with red dots indicating median values. The dashed red line indicates an adjusted P value of 0.05. The number above each violin indicates the percentage of edges with adjusted P values smaller than 0.05 in the corresponding edge commonality group. B. Heatmap of the normalized ranks for the 100 sampled edges with a commonality score of 43 across cell clusters. Each edge has a rank *r* in each cell cluster based on its absolute correlation value, from which a normalized rank r' = (r-1)/(R-1) is calculated, where R is the total number of edges in a cell network. A smaller normalized rank value indicates this edge is highly ranked among gene pairs and is encoded as dark blue colors in the heatmap. The row labels are cell cluster annotations and the column labels are gene pairs. The trees on both sides are generated via hierarchical clustering using the 'hclust' function in R with the "ward.D2" method on the Euclidean-based distances.

While we define the core network based on edges found in at least 43 cell clusters, it is possible that that these edges occur repeatedly but are limited to the same set of cell types, and so perhaps they may have more cell-specificity than we might expect for a hypothetical core network, whose structure should be nearly the same regardless of cell type. To explore this possibility, we clustered cell types based on the rank of 100 randomly sampled edges from edge commonality group 43 (**Fig. 3B**). This analysis revealed a group of about 20 cell clusters that contain nearly all of the core edges. The remaining cell clusters contained a variety of network edges along with highly ranked, but sub-threshold, network edges in their co-expression matrices. This indicates that the relative strength of the edges in the core varies by cell cluster.

Of all *Drosophila* organs, the brain may contain the greatest diversity of cell types that are currently resolved by scRNA-seq studies (Li et al. 2021). Having identified a network that is shared across cell types of the *Drosophila* brain, we then validated our analysis in an independent study of the same organ (Baker et al. 2021) (**Supplementary Material, section 2**), and extended our analysis to a study of cells from whole fly heads, which allowed us to survey an even wider range of cell types (Li et al. 2021) (**Supplementary Material, section 3)**. The independent study of the *Drosophila* brain included samples of adult flies of both sexes, treated with either sucrose, or sucrose supplemented with cocaine, and profiled by scRNA-seq (Baker et al. 2021). We focused our analysis on sucrose samples for both sexes, and found 1,738 commonly expressed genes across 47 cell clusters. Following network construction as above, we identified a network of 323 edges and 88 genes (**Fig. S8**). As in the analysis described above, in this brain core network, we did not identify edges found in all cell types. However, edges found in most cell types ranked higher than expected by chance in the remaining cell types (**Fig. S8**).

282 Turning to the analysis of the entire fly head, among the 76 cell clusters identified, we
283 found 842 commonly expressed genes, among which we identified a network of 21 edges and 20
284 genes (**Fig. S9**). Compared to the fly brain data, the fly head data included at least nine additional
285 cell clusters, which may represent cell types outside of the brain, and opens the possibility of
286 evaluating the persistence of the putative core network across a wider range of cell types. Here
287 again, the head core network did not have edges in all cells, but those expressed in most cells
288 ranked highly in the remaining cell types (**Fig. S9**). We next asked if there was a set of genes
289 observed commonly across all three network analyses. In fact, we found 62 genes that were shared
290 by the core network from fly brain atlas data and the second brain study (dataset of Baker et al.
291 2021) and seven genes shared with the fly head study (dataset of Li et al. 2021). Compared to
292 random samples from the commonly expressed genes in each data set, these overlaps among the
293 real networks were highly significant (empirical P value = 0.00099, calculated as the proportion
294 of random samples whose overlapped gene numbers were larger than the observed one, **Fig. S8H**
295 **and S9H**).

296

297 **Topological, functional and evolutionary signatures of the core cellular network**

298 Having defined a core cellular network, we next examined its topological, functional and
299 evolutionary properties. To evaluate the topological properties of the core network, we calculated
300 its clustering coefficient, and compared it to an ensemble of coefficients from pseudo-core
301 networks each with the same number of genes, edges, and degree distribution as the observed one.

302 A high clustering coefficient indicates that nodes of a network are densely connected to each other.
303 The observed clustering coefficient value, which has a range of 0 to 1, was 0.68, much higher than
304 the mean simulated value of 0.13 (range 0.11 - 0.16) (**Fig. S10**), suggesting the core network has
305 a dense connection structure.

306 Many complex gene networks can be divided into modules, where genes are more highly
307 interconnected within modules than between modules (Newman 2003). Modules identified from
308 gene co-expression networks tend to take part in the same biological processes or pathways
309 (Ruprecht et al. 2017; Wolfe et al. 2005). To explore the structural and functional organization of
310 this core, we decomposed it into highly connected modules using the Markov Clustering
311 Algorithm (**Methods**). In total, we identified four modules, each with at least five gene members
312 (**Fig. 4A and 4B; Table S2**). Most gene pairs within modules were positively correlated, while
313 the between-module connections tended to be negative (**Fig. S11)**. We then annotated each
314 module's biological function through Gene Ontology (GO) enrichment analysis. The results
315 revealed an array of housekeeping functions enriched within different modules (**Fig. 4C; Table**
316 **S3**), including pathways associated with ATP metabolism, glycolysis, translation, and proteostasis.
317 The largest module (module 3) contained 83 genes, and was enriched for ATP metabolic functions,
318 which suggests highly coordinated expression of genes involved in energy production. The second
319 largest module (module 2) contained 70 genes and was enriched for ribosome related functions,
320 such as cytoplasmic translation, suggesting tight correlation of expression among genes encoding
321 ribosomal proteins across cells. We note that this analysis was limited to mRNA, and so this
322 module does not contain the stoichiometrically synthesized rRNAs. Module 4 had 29 genes
323 associated with glycolysis, a process central to cellular energy homeostasis. Module 1 formed a

fully interconnected subnetwork without any edges to other modules (**Fig. 4B)**, and its gene members were highly enriched for heat-shock proteins (HSP) or co-chaperones, key players in protein folding.

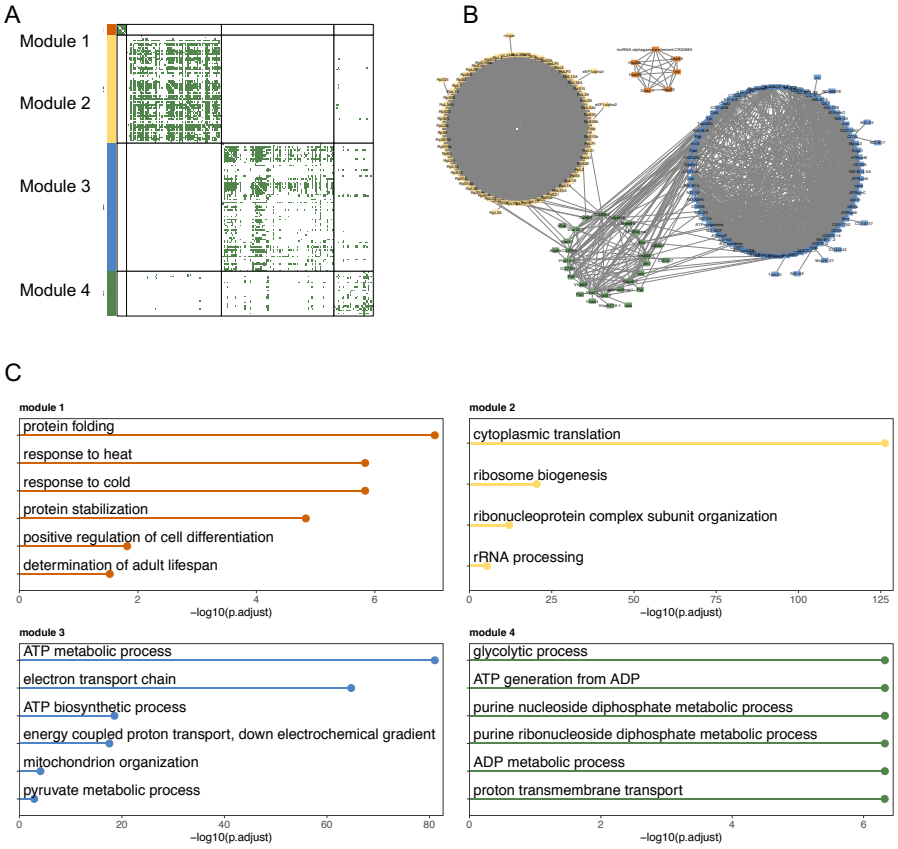

Figure 4. Gene modules in the defined core network. A. Heatmap of gene co-expression relationships and decomposed modules in the defined core cellular network. Modules that have at least five gene members are highlighted in different colors and numerically indexed. B. Network visualization of the core modules. C. Enriched GO terms for each core module. We used the R package 'clusterProfiler' to perform gene set enrichment analysis of Gene Ontology with a Bonferroni correction and an adjusted P value cutoff of 0.05. In each module, the top terms are shown (up to 6). A full list of enriched GO terms for each module is provided in **Table S3**.

334

335 To characterize the evolutionary signature of each module, we used phylostratigraphy,
336 assigning each gene in each module to one of 10 different evolutionary time periods (Domazet-
337 Lošo et al. 2017) (**Table S4**). Broadly speaking, genes in the core network were enriched for genes
338 with ancient origins, compared with other commonly expressed genes in the fly brain (**Fig. S12A**).
339 Looking in detail, the four modules of the core network differed significantly with respect to the
340 evolutionary age of their constituent genes (**Fig. S12B**). Gene members of the ribosomal (module
341 2) and protein folding (module 1) modules predated the divergence of the eukaryota, while those
342 of the ATP metabolic (module 3) and glycolysis (module 4) modules included genes distributed
343 across both ancient and relatively recent evolutionary time periods. While this analysis measures
344 the age of the gene members at the network nodes, rather than the age of network edges, these
345 diverse age signatures of different core modules suggest that some modules arose by integration
346 of both young and old genes, perhaps involving step-wise recruitment of young genes into
347 ancestral core modules.

348

349 **Discussion**

350 To what extent do all cells in an organism rely on a common core of interacting genes?
351 Would such a network be detectable among the expression patterns of diverse cell types? To
352 investigate these questions, we examined cell type-specific gene co-expression networks using fly

353 brain scRNA-seq and fly head snRNA-seq data. We describe an approach to find shared
354 components among multiple co-expression networks, identifying a small and highly clustered core
355 network. We validate this result, finding structurally and functionally similar networks in two
356 independent datasets, one from the fly brain, and a second from fly heads, the latter of which
357 includes a greater diversity of cell types. The core network we identify is composed of four co-
358 expression modules, each of which has distinct functional enrichment and distribution of gene age.

359 While the search for core networks is not new (Almaas et al. 2005; Neph et al. 2012;
360 Skinnider et al. 2021), our study is distinct in at least three ways from previous work. First, instead
361 of relying on consistent expression levels of individual genes to identify genes common across cell
362 types, we examined covariation between genes as the measure of functional commonality. This
363 focus provides not only a stricter criterion to infer gene function (Hughes et al. 2000), but also has
364 the potential to capture conserved gene regulatory networks (Yu et al. 2003; Stuart et al. 2003;
365 Segal et al. 2003). Second, a large body of studies has relied heavily on PPI data to derive
366 biological networks and find commonalities (Liu et al. 2020; Skinnider et al. 2021; Huttlin et al.
367 2021). While these studies are informative, they suffer from the bias that PPI data often lack
368 information on the degree of cell specificity of such interactions, and are enriched for highly-
369 studied proteins, which may give an incomplete picture of network structure (Skinnider et al. 2018;
370 Gillis et al. 2014; Schaefer et al. 2015). In our study, we analyzed transcriptome data, which
371 interrogate almost all genes in the genome and are less biased with respect to knowledge from
372 prior databases or existing literature. Third, we identified covarying gene pairs using scRNA-seq
373 data, which unlike bulk transcriptome data, or broad-scale PPI data, can be defined by cell type,
374 even within a single biological sample. In contrast to bulk transcriptomic analysis and PPI data,

375 where the cellular specificity of each interaction is largely ambiguous, scRNA-seq enabled us to
376 build cell type-specific networks at a resolution that was previously impossible.

377

378 **Topological properties of the core network**

379 With our current parameter choices, the defined core network is remarkably small when
380 compared to the much larger network of cell-specific interactions. In particular, in the fly brain,
381 the relative core size, calculated as the 2,140 core edges divided by 108,942, the number of edges
382 in a cell cluster-specific network, is only 2%. This number is relatively small compared with
383 previous studies of different biological networks. For example, Skinnider et al. (2021) constructed
384 tissue-specific PPI networks for seven mouse tissues and found that those universal PPIs shared
385 by all tissues typically occupy 3.7% of the total number of PPIs in a tissue (Skinnider et al. 2021).
386 Neph et al. (2012) built TF interaction networks for 41 cell types in humans and found that 5% of
387 interacting TFs were common to all cell types (Neph et al. 2012). Almaas et al. (2005) used flux-
388 balance analysis to study active metabolic reactions of *Escherichia coli* in 30,000 diverse
389 simulated environments and predicted that 90 of 758 (11.9%) reactions were always active
390 (Almaas et al. 2005).

391 Another prominent feature of this core network is its dense connectivity. The gene network
392 architecture we observed, which embodies extensive cell-type specific interactions along with a
393 shared and densely connected core, echoes findings from other types of biological networks. For

394 example, Liu et al. (2020) identified 13,764 PPIs in yeast across nine environments and found that
395 60% of PPIs were found in only one environment (Liu et al. 2020). They also show that PPIs that
396 were present in eight or more environments formed 'tight' modules of high node degree, while PPIs
397 present in three or fewer environments formed less-connected modules of smaller node degree.
398 Protein interaction networks based on just two human cell lines revealed that shared interactions
399 tend to reside in dense subnetworks and correspond to known protein complexes such as the
400 exosome and the COP9 signalosome (Huttlin et al. 2021). Similarly, network analyses of gene co-
401 expression from bulk transcriptomics in *Arabidopsis* or in humans suggest a highly connected
402 core, which appears alongside an extensive number of condition-specific gene interactions (He
403 and Maslov 2016; Lee et al. 2004). Taken together, these results suggest a universal organizing
404 principle in biological systems, where widely shared components of interaction networks are
405 relatively small and densely connected (Milo et al. 2002, 2004; Csermely et al. 2013).

406

407 **Is there a core network active in all cells?**

408 To determine whether there is a core co-expression network common to all cells, we
409 focused our analysis on genes that were expressed across all cell clusters. This enabled us to
410 directly compare the diverse co-expression patterns among a common set of genes. Single-cell
411 sequencing data suffer from a high level of sparsity. It has been estimated that the proportion of
412 zeros in bulk RNA-seq data is usually 10% - 40%, while in scRNA-seq data it can be as high as
413 90% (Jiang et al. 2022). Such sparsity may be due to both technical reasons, such as the lack of

414 efficiency in capturing the single-cell transcriptome during library preparation, and biological
415 reasons, such as the stochastic nature of gene expression (Sarkar and Stephens 2021). Data
416 sparsity, as a universal analytical challenge for scRNA-seq data, is an active research area, and
417 little consensus has been reached on how to handle this challenge (Lähnemann et al. 2020). Our
418 approach involved deriving cell cluster-specific co-expression matrices and only then did we look
419 at covariation of each gene pair across cell clusters, and we found that the patterns of co-expression
420 networks that we detected were robust to relatively high levels of sparsity (**Supplementary**
421 **Material, section 1**).

422 Even among the genes expressed in all cell clusters of the fly brain, we found that most
423 gene co-expression was cell cluster-specific, and yet there also existed a significantly large number
424 of genes whose co-expression occurred among multiple cell clusters. This enrichment of shared
425 co-expressed genes suggests the existence of a core co-expression network. While we identified
426 edges common to many cell clusters, there was no edge that was shared by all cell clusters. We
427 consider two alternative explanations for this observation.

428 First, it is possible that a common core co-expression network for all brain cells does not
429 exist, and that cell co-expression networks are so diverse as to lack such rigid network structure
430 across the cells of the *Drosophila* brain. Alternatively, while sparsity does not appear to present a
431 problem for our analysis, there is still the possibility that a core does exist, but that the statistical
432 inference of gene co-expression and the threshold values we use to build the network might not
433 fully resolve the core in all cells. To define the core, we specified multiple parameters, including
434 a gene correlation metric, a correlation threshold to select co-expressed genes, and an edge

435 commonality cutoff to extract a core network. Given that our definition of gene interactions is
436 based on statistical inference of correlations, we might simply fail to observe a real interaction in
437 one or more cell clusters due to type II error (false negatives). As we explored this parameter space,
438 moving from stringent to relaxed parameter values, a few features of the network became apparent.
439 In particular, more relaxed parameter values revealed co-expressed gene pairs observed in all cell
440 clusters in the brain, but this inherently increased the risk of type I error (false positive gene pairs),
441 as revealed by network permutation. While the network that we define lacks edges shared by all
442 cell clusters, edges that are present in most clusters tend to be highly ranked in those cell types
443 where they do not reach the threshold of significance. Thus, their absence from the top 5% of gene
444 pairs in some cells may be partly explained by their relatively weak co-expression strengths
445 compared to the cell cluster-specific gene pairs. Together, our analysis indicates the existence of
446 a core cellular network, though the size and composition that we define is conditional on parameter
447 choices.

448 It is likely that other factors not examined here also contribute to the size, structure and/or
449 function of a core network. These could include cell/organ type, sex, genotype, age, and so forth.
450 For example, one recent study using bulk RNA-seq data to examine gene co-expression network
451 dynamics with age in bats identified a small core network whose size decreases with age (Bernard
452 et al. 2022). That said, we find that the basic structure of our defined core network occurs in both
453 sexes when measured in the fly brain, a complex organ with highly heterogeneous cell type
454 composition. Projecting such an analysis to more organs, or even to a whole fly, would very likely
455 reveal an even smaller core, as the inclusion of a larger set of diverse cells would lead to a smaller
456 set of universal genes, and universal edges. Indeed, in our analysis of the fly head, which contains

457 more cell types than the fly brain, the uncovered core network was indeed smaller, yet still
458 significantly intersecting in gene identity with the core network from a unique sample of fly brains.
459 We therefore speculate that the relative core size that we observe in the fly brain data, which
460 includes 2% of cell cluster-specific network edges, might be an overestimate of the true core
461 network size for all cells in *Drosophila*. Future studies should apply our approach on an organism-
462 wide scale, in search of a small but ubiquitous core network testing found in all cells in an
463 organism.

464

465 **Limitation and future directions**

466 In this study, we sought a core of interacting genes found across cell clusters in the fly
467 brain. While the work described here benefits from access to high quality single-cell transcriptome
468 data, there are still several caveats worth noting. First, the fly brain cell atlas (Davie et al. 2018)
469 was generated using a mixture of two genotypes, and with cells from individuals of several ages.
470 Thus, genotype or age-specific gene co-expression patterns are yet to be revealed. Future studies
471 targeting individual genotypes and/or specific age groups might uncover a dynamic picture of how
472 core networks change over age or genotype conditions. Second, we inferred co-expressed gene
473 pairs from gene expression data statistically. Gene co-expression is not proof of gene co-
474 regulation, which may be more indicative of functional relationships. Further experimental work
475 is needed to validate the functional implications of these gene pairs. Lastly, we caution that the
476 analysis on fly brain samples is based on scRNA-seq data, and the fly head samples were processed

477 using single nuclear RNA-seq. These two approaches might capture different aspects of cell
478 activity (Wu et al. 2019; Denisenko et al. 2020; Thrupp et al. 2020).

479 Our approach, which is generic and interpretable, can be extended to incorporate more
480 data, such as additional tissues, species, or biological conditions in future studies. Although current
481 single cell techniques yield data with high levels of sparsity, somewhat ambiguous cell type
482 resolution, and other challenges, we anticipate that similar and more complete data are on the
483 horizon. We expect that more comprehensive knowledge of gene expression in all cells may reveal
484 that a core co-expression network, shared by all cells under all conditions, may be very limited in
485 size, perhaps even non-existent. This possibility does not rule out the idea that all cells share
486 common functions that are required for cell survival, but that these functions are not always
487 dependent on gene co-expression, or that the co-expression within such a network is weak relative
488 to the many conditional co-expression relationships that occupy cellular networks.

489

490 **Methods**

491 **Dataset collection and preprocessing**

492 We downloaded the fly brain atlas data from NCBI Gene Expression Omnibus (GEO
493 GSE107451). The original dataset contains expression data for 17,473 genes in 56,902 high-
494 quality brain cells grouped into 116 cell clusters. As a quality control step, we first removed 668

495 cells in a cell cluster named 'Hsp', as they represent stressed cells (Jasper Janssens, personal
496 communication). We then removed cells that had fewer than 200 expressed genes, fewer than 500
497 total unique molecular identifier counts, or a total fraction of mitochondrial gene expression
498 exceeding 30%. These criteria led to the removal of another 42 cells, leaving 56,192 cells. These
499 cells were assigned to 115 cell clusters from which we selected 37 cell clusters in females and 31
500 cell clusters in males that had at least 200 cells each.

501 We filtered genes for each cell cluster in each sex individually by removing genes that were
502 expressed in less than 15 cells, or in fewer than 0.5% of cells in that cell type. This gene filtering
503 procedure led to 7,795 genes as expressed in at least one cell type, 2,088 of which were commonly
504 expressed in all 67 cell clusters (**Fig. S1**).

505

506 **Constructing cell type-specific gene co-expression networks**

507 We used the bigScale2 algorithm (Iacono et al. 2019) to compute a gene-gene correlation
508 matrix for each cell cluster in each sex. This algorithm was tailored to mitigate the impact of sparse
509 counts at the single-cell level. It first groups cells into homogenous cell clusters, then performs
510 differential expression (DE) analysis between all pairs of clusters. With N clusters, we obtain
511 N*(N-1)/2 unique comparisons, and each comparison generates one z-score for each gene,
512 indicating the likelihood of an expression change between the corresponding two clusters. Finally,
513 bigScale2 uses transformed z-scores instead of original expression values to calculate Pearson

514 correlation coefficients. This z-score transformation allows us to detect correlations that would
515 otherwise be missed by drop-out events and other technical artifacts. Example scatter plots of z-
516 scores of gene pairs within cell clusters are presented in **Fig. S2.** In **Section 1** of the
517 **Supplementary Material,** we further demonstrate the ability of this algorithm to provide robust
518 estimates even with sparse data. To select highly correlated gene pairs for inclusion in a gene co-
519 expression network, we employed a signal-to-noise score approach and calculated this score across
520 various top percentile-based threshold values in each cell cluster separately. The highest signal-to-
521 noise score frequently occurs at a threshold value taking the top 5% of edges across cell clusters
522 (**Fig. S3 and S4**). Thus, we ranked gene pairs by their absolute correlation values in each cell
523 cluster separately, and placed the top 5% of correlated gene pairs into a cell cluster-specific co-
524 expression network, with the corresponding absolute correlation values ranging from r=0.39 to
525 0.85.

526

527 **Evaluating gene and edge commonality distributions**

528 To evaluate commonality and specificity across cell cluster-specific networks, we plot the node
529 and edge commonality distributions. The commonality of a node (gene) refers to the number of
530 cell clusters in which this gene is found to be co-expressed (edge) with at least one other gene. The
531 commonality of an edge linking a given pair of genes refers to the number of cell clusters in which
532 that specific edge is detected. We derived a mathematical approximation for the probability of a
533 gene pair to be co-expressed in a given number of cell types. As we focused on 2,088 commonly

534 expressed genes and selected the top 5% of highly correlated genes in each cell type, in randomized
535 data, a gene pair would have a probability $p = 0.05$ as being co-expressed in any one cell type.
536 Examining 67 cell clusters and using the binomial distribution, the probability $P(k)$ of a gene pair
537 being co-expressed in k cell types would equal

538 $$P(k) = C(67, k) * 0.05^{k} * (1 - 0.05)^{67-k}$$

539 Where $C(67, k)$ is the combinatorial number describing the number of ways of picking k items
540 from a pool of 67 cell clusters (that is, 67 choose k, or $\frac{67!}{k!(67-k)!}$). Following this equation, the
541 probability of a gene pair not being co-expressed in any cell type is 0.0321 at $k = 0$. Given the
542 2,088 commonly expressed genes, 2,108,888 non-redundant gene pairs were expected to co-
543 express in at least one cell type.

544

545 **Network randomization**

546 To obtain a null distribution for edge commonality distributions, we used a network
547 randomization approach. We randomized the edges in each cell type-specific network individually,
548 keeping the gene connectivities fixed using the rewire function from the R package iGraph (Csardi
549 and Nepusz 2006). A set of randomizations for all 67 cell clusters resulted in one pseudo edge

550 commonality distribution. We performed the randomization procedure 100 times and used the
551 ensemble of the 100 pseudo edge commonality distributions as the null distribution.

552

553 **Rank aggregation analysis**

554 Each gene pair or edge has a rank based on its absolute correlation value in a given cell
555 cluster. To determine if one edge is ranked consistently high across a set of cell clusters based on
556 its absolute correlation value, we used the aggregateRanks function from the R package
557 RobustRankAggreg (Kolde et al. 2012). This function is based on a probabilistic model of order
558 statistics and computes a derived P value for each edge. The derived P value ranges from 0 to 1
559 and serves as an upper bound of the computationally expensive exact P value, with a small value
560 indicating one edge is ranked consistently higher across cell types. We chose edge commonality
561 groups covering the full range of edge commonality scores, and within each edge commonality
562 group, we randomly sampled 100 edges. For each edge, we first collected the cell clusters in which
563 that edge was absent (i.e., was not in the top 5% of correlations), and then calculated the derived
564 P values from these cell clusters. The derived P values were corrected for multiple testing using
565 the p.adjust function in R with the Benjamini-Hochberg method, referred to as adjusted P values
566 hereafter. We plotted the adjusted P value distribution of the 100 sampled edges for each edge
567 commonality group separately.

568

569 **Module decomposition and functional annotation of the core**

570 To decompose the core cellular network into highly connected modules, we used the mcl
571 function from the R package MCL (Jäger 2015), which implements a Markov cluster algorithm to
572 identify clusters in networks. After module detection, we performed Gene Ontology (GO)
573 enrichment analysis of genes in each module using the R package clusterProfiler (Wu et al. 2021),
574 with a Bonferroni correction and an adjusted P value cutoff of 0.05. Significant GO terms were
575 identified and refined to reduce redundant GO terms via the simplify method from the
576 clusterProfiler package.

577

578 **Assigning genes into evolutionary age groups**

579 We downloaded data from a previous study to assign genes into different evolutionary age
580 groups using a phylostratigraphy framework (Domazet-Lošo et al. 2017). This framework allows
581 us to date the evolutionary origination time of a gene by identifying its homologs across the tree
582 of life. There were 13,794 genes assigned to 12 age groups in the original publication, 2,002 of
583 which overlapped with the 2,088 expressed genes in this study, including 911 genes in the oldest
584 age group "CellLife", 641 in "Eukaryota", 87 in "Opisthokonta", 122 in "Metazoa", 32 in

585 “Eumetazoa”, 73 in “Bilateria”, 14 in “Protostomia”, 16 in “Arthropoda”, 13 in “Pancrustacea”,
586 43 in “Insecta”, 39 in “Diptera” and 11 in the youngest age group “Drosophila”.

587

588 **Data access**

589 The fly brain atlas dataset analyzed during the current study is available at GEO with accession
590 number GSE107451. The dataset from Baker et al. (2021) is available at GEO with accession
591 number GSE152495. The dataset from Li et al. (2021) is available from the Fly Cell Atlas website
592 (https://flycellatlas.org/). R scripts for data analyses are available from the following GitHub
593 repository: https://github.com/mingwhy/fly.brain.core_coexpr.net.

594

595 **Competing interests statement**

596 The authors declare no competing interests.

597

Acknowledgements

This work was supported in part by National Institute on Aging grants R21AG56872901, R01AG057330, and R01AG063371. This work was in part facilitated through the use of advanced computational, storage, and networking infrastructure provided by the Hyak supercomputer system at the University of Washington.

Author contributions： M. Y. and D.P. designed the study. M.Y. collected and analyzed the data. M.Y., B.H. and D.P. interpreted the data and wrote the manuscript.

References

Almaas E, Oltvai ZN, Barabási A-L. 2005. The activity reaction core and plasticity of metabolic networks. *PLoS computational biology* **1**: e68.

Baker BM, Mokashi SS, Shankar V, Hatfield JS, Hannah RC, Mackay TFC, Anholt RRH. 2021. The Drosophila brain on cocaine at single-cell resolution. *Genome research* **31**: 1927–1937.

Barabasi A-L, Oltvai ZN. 2004. Network biology: understanding the cell's functional organization. *Nature reviews genetics* **5**: 101–113.

Bernard G, Teulière J, Lopez P, Corel E, Lapointe F-J, Bapteste E. 2022. Aging at evolutionary crossroads: longitudinal gene co-expression network analyses of proximal and ultimate causes of aging in bats. *Molecular biology and evolution* **39**: msab302.

Csardi G, Nepusz T. 2006. The igraph software package for complex network research. *InterJournal, complex systems* **1695**: 1–9.

Csermely P, London A, Wu L-Y, Uzzi B. 2013. Structure and dynamics of core/periphery networks. *Journal of Complex Networks* **1**: 93–123.

Davie K, Janssens J, Koldere D, de Waegeneer M, Pech U, Kreft Ł, Aibar S, Makhzami S, Christiaens V, González-Blas CB. 2018. A single-cell transcriptome atlas of the aging Drosophila brain. *Cell* **174**: 982–998.

Denisenko E, Guo BB, Jones M, Hou R, de Kock L, Lassmann T, Poppe D, Clément O, Simmons RK, Lister R. 2020. Systematic assessment of tissue dissociation and storage biases in single-cell and single-nucleus RNA-seq workflows. *Genome biology* **21**: 1–25.

Domazet-Lošo T, Carvunis A-R, Albà M, Šestak MS, Bakarić R, Neme R, Tautz D. 2017. No evidence for phylostratigraphic bias impacting inferences on patterns of gene emergence and evolution. *Molecular biology and evolution* **34**: 843–856.

Farahbod M, Pavlidis P. 2020. Untangling the effects of cellular composition on coexpression analysis. *Genome research* **30**: 849–859.

Ghadie MA, Coulombe-Huntington J, Xia Y. 2018. Interactome evolution: Insights from genome-wide analyses of protein–protein interactions. *Current opinion in structural biology* **50**: 42–48.

Gillis J, Ballouz S, Pavlidis P. 2014. Bias tradeoffs in the creation and analysis of protein–protein interaction networks. *Journal of proteomics* **100**: 44–54.

Greene CS, Krishnan A, Wong AK, Ricciotti E, Zelaya RA, Himmelstein DS, Zhang R, Hartmann BM, Zaslavsky E, Sealfon SC. 2015. Understanding multicellular function and disease with human tissue-specific networks. *Nature genetics* **47**: 569–576.

Harris BD, Crow M, Fischer S, Gillis J. 2021. Single-cell co-expression analysis reveals that transcriptional modules are shared across cell types in the brain. *Cell Systems* **12**: 748–756.

Hart Y, Alon U. 2013. The utility of paradoxical components in biological circuits. *Molecular cell* **49**: 213–221.

He F, Maslov S. 2016. Pan-and core-network analysis of co-expression genes in a model plant. *Scientific reports* **6**: 1–11.

Hughes TR, Marton MJ, Jones AR, Roberts CJ, Stoughton R, Armour CD, Bennett HA, Coffey E, Dai H, He YD. 2000. Functional discovery via a compendium of expression profiles. *Cell* **102**: 109–126.

Huttlin EL, Bruckner RJ, Navarrete-Perea J, Cannon JR, Baltier K, Gebreab F, Gygi MP, Thornock A, Zarraga G, Tam S. 2021. Dual proteome-scale networks reveal cell-specific remodeling of the human interactome. *Cell* **184**: 3022–3040.

Iacono G, Massoni-Badosa R, Heyn H. 2019. Single-cell transcriptomics unveils gene regulatory network plasticity. *Genome biology* **20**: 1–20.

Jäger ML. 2015. MCL: Markov Cluster Algorithm. R package version 1.0. *https://CRANR-project.org/package=MCL*.

Jiang R, Sun T, Song D, Li JJ. 2022. Statistics or biology: the zero-inflation controversy about scRNA-seq data. *Genome Biology* **23**: 1–24.

656 Kolde R, Laur S, Adler P, Vilo J. 2012. Robust rank aggregation for gene list integration and meta-
657 analysis. *Bioinformatics* **28**: 573–580.
658 Lähnemann D, Köster J, Szczurek E, McCarthy DJ, Hicks SC, Robinson MD, Vallejos CA,
659 Campbell KR, Beerenwinkel N, Mahfouz A, et al. 2020. *Eleven grand challenges in single-cell*
660 *data science*. Genome Biology.
661 Lee HK, Hsu AK, Sajdak J, Qin J, Pavlidis P. 2004. Coexpression analysis of human genes across
662 many microarray data sets. *Genome research* **14**: 1085–1094.
663 Lehner B, Fraser AG. 2004. Protein domains enriched in mammalian tissue-specific or widely
664 expressed genes. *Trends in Genetics* **20**: 468–472.
665 Li H, Janssens J, de Waegeneer M, Kolluru SS, Davie K, Gardeux V, Saelens W, David F, Brbić
666 M, Leskovec J. 2021. Fly Cell Atlas: a single-cell transcriptomic atlas of the adult fruit fly. *bioRxiv*.
667 Lim WA, Lee CM, Tang C. 2013. Design principles of regulatory networks: searching for the
668 molecular algorithms of the cell. *Molecular cell* **49**: 202–212.
669 Liu Z, Miller D, Li F, Liu X, Levy SF. 2020. A large accessory protein interactome is rewired across
670 environments. *Elife* **9**: e62365.
671 Milo R, Itzkovitz S, Kashtan N, Levitt R, Shen-Orr S, Ayzenshtat I, Sheffer M, Alon U. 2004.
672 Superfamilies of evolved and designed networks. *Science* **303**: 1538–1542.
673 Milo R, Shen-Orr S, Itzkovitz S, Kashtan N, Chklovskii D, Alon U. 2002. Network motifs: simple
674 building blocks of complex networks. *Science* **298**: 824–827.
675 Mitra K, Carvunis A-R, Ramesh SK, Ideker T. 2013. Integrative approaches for finding modular
676 structure in biological networks. *Nature Reviews Genetics* **14**: 719–732.
677 Neph S, Stergachis AB, Reynolds A, Sandstrom R, Borenstein E, Stamatoyannopoulos JA. 2012.
678 Circuitry and dynamics of human transcription factor regulatory networks. *Cell* **150**: 1274–1286.
679 Newman MEJ. 2003. The structure and function of complex networks. *SIAM review* **45**: 167–256.
680 Parfrey LW, Lahr DJG. 2013. Multicellularity arose several times in the evolution of eukaryotes
681 (Response to DOI 10.1002/bies. 201100187). *Bioessays* **35**: 339–347.
682 Promislow D. 2005. A regulatory network analysis of phenotypic plasticity in yeast. *The American*
683 *naturalist* **165**: 515–523.
684 Proulx SR, Promislow DEL, Phillips PC. 2005. Network thinking in ecology and evolution. *Trends*
685 *in ecology & evolution* **20**: 345–353.
686 Ruprecht C, Proost S, Hernandez-Coronado M, Ortiz-Ramirez C, Lang D, Rensing SA, Becker
687 JD, Vandepoele K, Mutwil M. 2017. Phylogenomic analysis of gene co-expression networks
688 reveals the evolution of functional modules. *The Plant Journal* **90**: 447–465.
689 Sarkar A, Stephens M. 2021. Separating measurement and expression models clarifies confusion
690 in single-cell RNA sequencing analysis. *Nature genetics* **53**: 770–777.
691 Schaefer MH, Serrano L, Andrade-Navarro MA. 2015. Correcting for the study bias associated
692 with protein–protein interaction measurements reveals differences between protein degree
693 distributions from different cancer types. *Frontiers in genetics* **6**: 260.
694 Segal E, Shapira M, Regev A, Pe'er D, Botstein D, Koller D, Friedman N. 2003. Module networks:
695 identifying regulatory modules and their condition-specific regulators from gene expression data.
696 *Nature genetics* **34**: 166–176.
697 Skinnider MA, Scott NE, Prudova A, Kerr CH, Stoynov N, Stacey RG, Chan QWT, Rattray D,
698 Gsponer J, Foster LJ. 2021. An atlas of protein-protein interactions across mouse tissues. *Cell*
699 **184**: 4073–4089.
700 Skinnider MA, Stacey RG, Foster LJ. 2018. Genomic data integration systematically biases
701 interactome mapping. *PLoS computational biology* **14**: e1006474.
702 Sonawane AR, Platig J, Fagny M, Chen C-Y, Paulson JN, Lopes-Ramos CM, DeMeo DL,
703 Quackenbush J, Glass K, Kuijjer ML. 2017. Understanding tissue-specific gene regulation. *Cell*
704 *reports* **21**: 1077–1088.
705 Sorrells TR, Johnson AD. 2015. Making sense of transcription networks. *Cell* **161**: 714–723.

706 Stuart JM, Segal E, Koller D, Kim SK. 2003. A gene-coexpression network for global discovery of
707 conserved genetic modules. *science* **302**: 249–255.
708 Tanay A, Regev A. 2017. Scaling single-cell genomics from phenomenology to mechanism.
709 *Nature* **541**: 331–338.
710 Thompson D, Regev A, Roy S. 2015. Comparative analysis of gene regulatory networks: from
711 network reconstruction to evolution. *Annual review of cell and developmental biology* **31**: 399–
712 428.
713 Thrupp N, Frigerio CS, Wolfs L, Skene NG, Fattorelli N, Poovathingal S, Fourne Y, Matthews PM,
714 Theys T, Mancuso R. 2020. Single-nucleus RNA-seq is not suitable for detection of microglial
715 activation genes in humans. *Cell reports* **32**: 108189.
716 Trapnell C. 2015. Defining cell types and states with single-cell genomics. *Genome research* **25**:
717 1491–1498.
718 Wagner A. 2012. Metabolic networks and their evolution. In *Evolutionary systems biology*, pp.
719 29–52, Springer.
720 Wolfe CJ, Kohane IS, Butte AJ. 2005. Systematic survey reveals general applicability of" guilt-by-
721 association" within gene coexpression networks. *BMC bioinformatics* **6**: 1–10.
722 Wu H, Kirita Y, Donnelly EL, Humphreys BD. 2019. Advantages of single-nucleus over single-cell
723 RNA sequencing of adult kidney: rare cell types and novel cell states revealed in fibrosis. *Journal*
724 *of the American Society of Nephrology* **30**: 23–32.
725 Wu T, Hu E, Xu S, Chen M, Guo P, Dai Z, Feng T, Zhou L, Tang W, Zhan L. 2021. clusterProfiler
726 4.0: A universal enrichment tool for interpreting omics data. *The Innovation* **2**: 100141.
727 Yu H, Luscombe NM, Qian J, Gerstein M. 2003. Genomic analysis of gene expression
728 relationships in transcriptional regulatory networks. *Trends in Genetics* **19**: 422–427.
729 Zhang L, Li W-H. 2004. Mammalian housekeeping genes evolve more slowly than tissue-specific
730 genes. *Molecular biology and evolution* **21**: 236–239.
731 Zhu J, He F, Hu S, Yu J. 2008. On the nature of human housekeeping genes. *Trends in genetics*
732 **24**: 481–484.